Center for Big Data in Translational Genomics

The University of California Santa Cruz

PIs: David H. Haussler, David Patterson, and Laura Van't Veer Grant Number: 1-U54HG007990-01

The Center for Big Data in Translational Genomics is a multi-institution partnership coordinated by the University of California at Santa Cruz to create scalable infrastructure for the broad application of genomics in biomedicine. Our U.S. partners include UC San Francisco, UC Berkeley, Oregon Health Science University, Caltech, and several major big data companies. International partners include the European Bioinformatics Institute, the Sanger Centre, the Ontario Institute for Cancer Research and a computer systems provider. The Center will make software solutions interoperable through the development of standard Application Programming Interfaces (APIs) and tools at multiple levels, from raw sequence data to genetic variation and functional data, through to systems, pathways and phenotypes. The overriding goal is to create implementations capable of handling genomics datasets that are orders of magnitude larger than those that can now be handled. The APIs and all academic reference implementations will be open source, while several major corporate partners not funded by the project will provide proprietary implementations, creating a competitive ecosystem of interoperable big data genomics software. All-comers extensive benchmarking will be performed on all implementations within and external to our center to identify best-of-breed and results made broadly available. Design will be in part driven by the needs of a diverse set of separately funded specific biomedical projects that will serve as pilots. These include the Pan-Cancer whole genome analysis project of the International Cancer Genomics Consortium to analyze 2,000 cancer genomes, the UK10K project to analyze 10,000 personal genomes, the UCSF-led I-SPY2 adaptive breast cancer trial, and the omics-guided leukemia therapy project BeatAML at Oregon Health Sciences University. PUBLIC HEALTH RELEVANCE: At least half of all diseases have a substantial genomic component, often including contributions from the millions of individually rare but collectively common genetic variations. Only by studying the genomes and transcriptomes of very large numbers of individuals will scientists have the statistical power to discover and understand this vital aspect of the genomic contribution to disease. For this it is essential that genomics is brought into the big data era, so that analyses of huge datasets is possible and precision diagnosis and treatment based on genomic information is widely deployed.